Expanding Small Corpora to Aid People with Communication Impairments

Gyula Vörös

There are people with various movement and cognitive disorders who are unable to speak or write. Many of them are able to communicate with pictorial symbols, but this still narrows the circle of possible communication partners to those who are familiar with this method.

The effects of this problem can be diminished by translating symbol sequences to sentences in natural language. For example, a sequence of symbols that stand for *'I'*, *'have'* and *'sandwich'*, should be translated to something like *'I would like to have a sandwich'*. An approach that is satisfactory in practice is to define all possible sentences manually, for example, by storing them in a text corpus. The construction of a such a corpus would require a lot a work.

In this paper, it is shown that writing all the sentences is not necessary; only a small initial seed corpus is required, which can be expanded automatically. A method is proposed for this expansion, based on replacing the nouns of the input sentences with other nouns. The resulting candidate sentences are then filtered using n-gram statistics from a much larger corpus. This is similar to the 5gramSum method described in [1]. The sentences in the expanded corpus will be similar to the initial ones in structure, but different in content. The method is language-independent, although using it in agglutinative languages such as Hungarian would require morphological processing.

The error rate of the proposed method was evaluated on a small corpus, which was collected from an English language learning website, and contained dialogues. The size of the corpus was doubled by introducing new sentences using our method. For example, the sentence *'I would like to buy some bread'* was generated from the original *'I would like to buy some beef'*. Samples of 100-100 sentences were randomly selected from the original and the new corpora. Additionally, as a baseline, 100 sentences were generated from the original corpus by replacing nouns randomly, without filtering them using n-gram statistics. Two annotators evaluated each sentence from the samples. The results show that the majority of the introduced sentences were potentially useful. The method produced 3-4 times as many good sentences as the baseline. This indicates that it may be possible to define a small corpus, extend it automatically, and use the resulting set of sentences for communication.

The practical applicability of the method was demonstrated by implementing a sentence production prototype software for alternative communication. A symbol set was defined that enables communication in a food buying situation. A small corpus of appropriate sentences was defined manually, then expanded automatically by including other words from the symbol set. The system was able to produce new meaningful sentences. This way, the amount of manual work necessary to create a communication aid was reduced considerably.

Acknowledgements

This work was carried out as part of the EITKIC 12-1-2012-0001 project, which is supported by the Hungarian Government, managed by the National Development Agency, financed by the Research and Technology Innovation Fund and was performed in cooperation with the EIT ICT Labs Budapest Associate Partner Group.

I would like to thank the work of the people involved with the project, especially András Lőrincz, András Sárkány, Anita Verő, Balázs Pintér and Brigitta Miksztai-Réthey.

References

[1] Sinha, R. and Mihalcea, R.: Combining lexical resources for contextual synonym expansion, in *Proceedings of the International Conference RANLP*, pp. 404–410 (2009).